编者按:用户画像是真实用户的虚拟代表,是建立在一系列真实数据上的目标用户模型。构建用户画像的过程是了解用户的过程,通过将具象的信息数据抽象为用户特征,可以精准地定位目标群体,预测用户的真实需求和潜在需求,为个性化服务、推荐系统、精准营销等带来巨大的应用价值。本组关于用户画像的论文,分别针对学术博客、网络群体、微信用户、图书馆用户进行用户画像模型构建研究,可为用户画像领域相关研究提供借鉴。

学术博客用户画像模型构建与实证*

——以科学网博客为例

袁润[1] 王琦[2]
[1] 江苏大学图书馆 镇江 212013 [2] 江苏大学科技信息研究所 镇江 212013

摘要:［目的/意义］用户画像理论可用于标记学术群体的行为特征,为精准识别用户、服务学术型社交平台的精准营销、改善冷启动时期用户体验提供依据和参考。［方法/过程］利用Python和R语言编写获取和处理公开用户行为数据的程序,从博客的基本属性、积极性、权威性、博文影响力、兴趣偏好等5个维度构建用户画像概念模型,以科学网博客用户行为数据为例,开展实证研究。［结果/结论］提出刻画学术博客用户特征的具体指标和计算方法,表明用户画像模型对学术社交平台的管理和运营具有一定的理论意义和应用价值。

关键词: 学术博客 用户画像 R语言 案例分析
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2019.22.002

1 引言

交互设计之父A. Cooper[1]最早提出用户画像(User Profile)的概念,用户画像是真实用户的虚拟代表,是建立在一系列真实数据上的目标用户模型。真实数据主要指用户信息数据,包括静态数据(相对稳定的用户属性数据)和动态数据(不断变化的用户行为数据)两个部分。用户画像是基于用户属性和用户行为抽取出一个或一类用户的标签,对用户信息进行结构化描述。构建用户画像的过程也是了解用户的过程,通过将具象的信息数据抽象为用户特征,可以精准地定位目标群体,预测用户的真实需求和潜在需求。目前,用户画像被广泛应用于精准营销[2]、智能推荐[3-5]、产品研发[6]等领域,多采用统计[7]、贝叶斯网络[8]、机器学习[9-11]、主题模型[12]、聚类分析[4]、层次分类[13]等方法构建用户画像。

现有研究中存在许多基于社交网络平台的用户画像研究。A. Raghuram等[14]提出了一种高效的监督机器学习方法,将Twitter用户分为6个兴趣类别。R. Jiamthapthaksin等[15]基于Naive Bayes、ANN和SVM构建用户兴趣特征模型,并利用Facebook数据集验证模型的有效性。韩梅花等[16]将用户画像与阅读疗法结合起来,通过分析微博文本,计算抑郁情感指数,得到用户画像。王凌霄等[17]以知乎为例,从用户资历、参与度、回答质量、发展趋势4个方面构建用户画像。刘海鸥等[18]深入挖掘QQ群、天涯论坛、人人网等社交平台并构建用户画像模型。崔超等[19]基于用户画像理论、知识与用户间的关系提出了知识社区用户画像数据采集和模型构建实现思路。余传明等[11]对股吧的

* 本文系国家社会科学基金项目"图书馆知识发现服务的功能定位和建设策略研究"(项目编号:14BTQ018)研究成果之一。
作者简介:袁润(ORCID:0000-0003-4428-874X),教授,硕士生导师;王琦(ORCID:0000-0003-2294-9965),硕士研究生,通讯作者,E-mail:569045749@qq.com。
收稿日期:2019-03-03 修回日期:2019-05-30 本文起止页码:13-20 本文责任编辑:易飞

用户发文内容进行深度学习，结合股吧用户的粉丝数量、影响力、关注量等行为特征，提出一种“行为－内容”融合模型，识别噪声投资者这一特定用户画像。陈天歌[20]从不同角度刻画微信用户性格画像和多个品牌画像，以本我、自我和超我为逻辑脉络，对各版本画像进行对比分析。周文静[21]将加权概念兴趣、加权关键字兴趣合并为用户兴趣维度画像，与用户基本属性维度画像一起组成最终的校园论坛用户画像。

社交网络平台按其是否专门用于学术交流，可分为学术型和非学术型社交平台两类。现有的用户画像研究集中于非学术型社交平台，对学术型社交平台关注较少。学术型社交平台的用户群体是对科研工作感兴趣的科研工作者，他们利用平台创建个人信息、发布科研成果、开展学术交流，实现知识的交流、传播与共享，体现了 Science2.0 开放、共享和协作等现代科学研究理念。本文以科学网博客为例，将获取的用户属性和行为数据归纳为博客基本属性、积极性、权威性、影响力和主题偏好等 5 个维度的指标，据此开展学术博客用户画像实证研究，对现有的研究进行补充和完善。用户画像的结果可用于平台的知识产品精准推送服务、用户识别，为学术交流平台建设与运营提供决策依据，具有一定的学术价值和现实意义。

2 学术博客用户画像模型

博客和博文是网络社交平台上的两类主要实体。任何学者都可以在科学网实名注册成为博客，通过发表博文参与学术交流。博客的所有网络行为，例如注册、发布、分类、标注、阅读、推荐、评论、下载、引用、访问、留言和建立好友关系等，都被平台记录下来，本文将该记录称之为用户行为数据。用户行为数据越丰富，越能精确刻画用户特征，但是受搜集成本、技术及隐私保护的限制，部分用户行为数据难以获取。为了开展实证研究，本文采集到 20 个数据项，涵盖了用户主要行为数据，如表 1 所示，其中 $a_1－a_4$ 是用户的基本情况（B），$a_5－a_{10}$反映博客使用平台的积极性（V），a_4、$a_{11}－a_{14}$反映用户在博客平台的权威性（Q），a_{15}、a_{17}、a_{18}则从博文阅读和博文互动两个方面反映了博文影响力情况（I），a_{19}、a_{20} 可以提取描述主题偏好的特征词组（G）。用户行为数据分为数值型和字符型两类，为了利用这些数据对学术博客画像，本文提出了 5 个维度的用户画像模型（UPM），用公式（1）表示如下：

$$UPM = \{B, V, Q, I, G\} \quad (1)$$

表 1 科学网博客用户行为数据项

编号	数据名称	注释
a_1	用户 ID	用户唯一标识符，为用户属性信息，可直接获取
a_2	姓名	用户唯一标识符对应的用户姓名，为用户属性信息，可直接获取
a_3	研究领域	用户注册时选择的学科分类和研究方向，为用户属性信息，可直接获取
a_4	头衔	涵盖用户的学历或者职称情况，具有等级性，为用户属性信息且能反映用户权威性，可直接获取
a_5	博文数	用户创作的博文数量之和，与积极性正相关
a_6	活跃度数	平台给予用户登录、分享、推广等行为的奖励，与积极性正相关，可直接获取
a_7	分享数	用户在博客平台的转发博文次数之和，与积极性正相关，可直接获取
a_8	主题数	用户在博客平台发布主题帖的次数，与积极性正相关，可直接获取
a_9	回帖数	用户在博客平台回复主题帖的次数，与积极性正相关，可直接获取
a_{10}	在线时长	用户在平台的累计使用时长，与积极性正相关，可直接获取
a_{11}	好友数	与用户建立好友关系的人数，与权威性正相关，可直接获取
a_{12}	主页访问数	博客主页被其他用户访问的人次总数，与权威性正相关，可直接获取
a_{13}	被推荐总数	用户发布的博文被其他用户推荐到平台首页的人次总数，与权威性正相关，从博文数据集中统计
a_{14}	精选博文数	用户创作博文中被遴选为的精选博文篇数，与权威性正相关，从博文数据集中统计
a_{15}	阅读数	用户的博文被其他用户阅读的次数，与博文影响力正相关，从博文数据集中统计
a_{16}	被评论总数	用户发布的博文被其他用户评论的总次数，从博文数据集中统计
a_{17}	被推荐数	用户博文被其他用户推荐到平台首页的次数，与博文影响力正相关，从博文数据集中统计
a_{18}	被评论数	用户博文被其他用户评论的次数，与博文影响力正相关，从博文数据集中统计
a_{19}	系统分类	用户为博文添加系统给定的分类标签，可用于检索和分类，可直接获取
a_{20}	个人分类	用户为博文提炼出的分类标签，可用于检索和分类，可直接获取

2.1 基本属性

本文将研究领域(a_3)视为博客基本属性,该参数能比较准确地描述其所属学科领域。此外,用户 id、学历、职称等客观参数是静态的或一段时间内相对稳定且公开的数据,利用平台公开渠道可以获取此类数据。博客用户在注册时所填写职称、学历等信息,本文统一归类到"头衔",即数据项 a_4。由于这类数据规范性不太好,本文对其作了等级量化处理,方法如表 2 所示。这样,博客基本属性可以表达如下:

$$B = \{EduPos, ResF\} \tag{2}$$

表 2　头衔信息等级量化处理

等级	头衔信息
0	"无"
1	"本科以下""本科"
2	"助理教授""助理研究员""助理编辑""研究生""硕士在读""硕士"
3	"讲师""编辑""博士在读"
4	"副教授""副研究员""副高""副主任""副编审""博士后""博士"
5	"院士""教授""研究员""编审"

2.2 积极性

积极性指标(V)与博文数、活跃度数、分享数、主题数、回帖数、在线时长等博客行为数据正向相关,本文将这 6 项数据的熵权值定义为博客积极性指标,用公式(3)表示:

$$V = \sum \omega \cdot \alpha \tag{3}$$

其中 α 表示行为数据项 $a_5 - a_{10}$ 的归一化值,ω 为其权重系数。归一化计算公式如下:

$$\alpha = a / a_{max} \tag{4}$$

熵权法利用各个数据项所提供信息的不确定性来确定各项权重,适用于各类赋值问题,计算行为数据项权重系数的方法如公式(5)所示:

$$\omega = \frac{1 - e}{\sum_{i=1}^{n}(1 - e)} \tag{5}$$

其中,e 为行为数据项信息熵,计算方法如公式(6)所示:

$$e = -\frac{1}{\log(n)}\left(\frac{\alpha}{\sum \alpha}\right)\log\left(\frac{\alpha}{\sum \alpha}\right) \tag{6}$$

这里公式(3)取指标 V 的前 25% 为高积极性群体(H),前 50% 为中等积极性群体(M),前 75% 为普通积极性群体(C),其余为低积极性群体(L)。阈值的设定由平台管理方确定,根据实际需求调整可获得不同积极性群体。

2.3 权威性

权威性主要受到用户自身的学术地位和博客内容权威性的影响。数据项 a4 包含的用户学历和职称信息可以反映用户的学术地位。博客内容权威性的衡量可通过博客内容传播覆盖度来表示,随着关注博客动态的人数增多,博文的传播速度也会随之变快,博客权威性也越大[22]。传播覆盖度与博客的好友数、主页访问数、精选博文数、被推荐总数等博客行为数据正向相关。本文将这 5 项数据的熵权值定义为博客权威性指标,用公式(7)表示:

$$Q = \sum \omega \cdot \beta \tag{7}$$

公式(7)可参照公式(4)－公式(6)处理数据项和确定权重系数,Q 指标阈值划分标准与 V 指标相同。

2.4 博文影响力

博文影响力可以量化,是以博文内容的形式改变其他用户思想和行为(阅读、推荐、评论等行为)的能力[22]。张晓阳等[23]、郑超等[24]扩大 h 指数的适用范围,基于博文阅读数、被评论数,综合考虑博文内容的质与量,评估学术博客影响力。本文在前人的基础上进一步完善博文影响力评估指标体系,从博文阅读数和博文互动数(被评论数和推荐数)两个视角量化博文影响力。根据 h 指数的推论,定义观测统计量博文阅读数(c)和博文互动数(q):

$$c = \sqrt{a_{15}} \tag{8}$$

$$q = \sqrt{(a_{17}^{2} + a_{18}^{2})} \tag{9}$$

h 指数的数学公式如下:

$$h_c = \max\{r_1 : r_1 \leqslant c\};\ h_q = \max\{r_2 : r_2 \leqslant q\} \tag{10}$$

其中 r_1 是观测量 c 降序排列的博文的序次,r_2 是观测量 q 降序排列的博文的序次。实际应用中发现,h 指数存在同值且取值水平较低的情况。由于学术博客的社交属性,用户行为数据存在稀疏性,上述现象更为显著。针对 h 指数的不足,金碧辉等[25]提出了 R 指数。R 指数是 h 核内论文总被引频次的平方根,其度量结果可以有效区分同值 h 指数且不改变 h 核的形态。R 指数的数学公式如下:

$$R = \sqrt{\sum_{j=1}^{h} c_j} \tag{11}$$

式中 c_j 表示 h 核内第 j 篇论文的被引频次,且 $c_j \geqslant h$。此处 c_j 表示 h 核内第 j 篇博文的阅读数或互动数且 $h \in \{h_c, h_q\}$。将 h 指数与 R 指数组合使用,可以有效弥补 h 指数的不足,更好地评估和区分博客博文影响力,用公式(12)表示:

$$I = \{(h_c, R_c), (h_q, R_q)\} \tag{12}$$

其中,h_c 的前 25% 为高阅读影响力群体(H),前 50% 为中等阅读影响力群体(M),前 75% 为普通阅读影响力群体(C),其余为低阅读影响力群体(L);h_q 的

前 25% 为高互动影响力群体(H),前 50% 为中等互动影响力群体(M),前 75% 为普通互动影响力群体(C),其余为低互动影响力群体(L)。

2.5 兴趣偏好

兴趣是人们活动的巨大动力,兴趣偏好是用户创作博文时做出的理性的、具有倾向性的选择,学术博客兴趣偏好反映了学者的学术兴趣方向,通常可以用若干主题词(关键词)来描述。科学网博客平台为博文设置了"系统分类"和"个人分类"两种分类途径,博客将其发表的博文归类到某一系统分类的同时,还可以采用个人分类进一步细分。如果将系统分类视为第一类节点,将个人分类视为另一类节点,则这两类节点刚好构成二分网络关系。

二分网络属于复杂网络的一种,通过构建二分网络提取特征词组可以描述产品特征[26]、用户兴趣爱好[27]。若给定无向图 $G = (V, E)$,对本文而言,顶点 V 分别为 $a_{19}(V_1)$ 和 $a_{20}(V_2)$。显然,$V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$ 且 $\forall e = (u, v) \in E$,均有 $u \in V_1, v \in V_2$,满足二分网络条件。

一般地,若将系统分类词组记为 V1,用户分类词组记为 V2,则二分网络记为:

$$G = (V_1, V_2, E) \tag{13}$$

本文采用 R 语言 bipartite 包,创建了博客的"系统分类 – 个人分类"二分网络,利用 computeModules 函数划分网络社区,按权重排序提取到分类词组,以此描述博客兴趣偏好。

3 学术博客用户画像实证分析

3.1 数据来源与获取

在注重用户隐私保护的前提下,本文通过 Python 语言编写程序采集学术博客用户行为数据。采集对象选择拥有精选博文的用户,用户 url 采集时间为 2018 年 12 月 12 日,采集到 3 799 条不重复用户 url。在采集博文数据之前,对原始 url 数据进行简单的人工处理,剔除 146 条因设置隐私权限等因素造成数据缺失的 url,博文数据项获取时间为 2018 年 12 月 19 日。

爬取数据过程中构建了 BlogUsers 和 BlogContents 两个原始数据集。采集完成后对采集到的数据进行必要的处理:BlogUsers 中与博文有关的数据以 BlogContents 实际获取数据为准,集中统计阅读总数、被推荐总数和被评论总数等数据项;博文数据中极端异常或大量数据缺失的用户予以剔除,最终得到 2 339 位有效用户数据和 437 832 条博文数据。采集到的 40 万条博文记录当中,有效评论次数累计达到 313 余万次,有效推荐次数累计达到 283 余万次,总阅读次数超过 14.29 亿次,可见科学网博客具有较大的影响力,对学术交流和传播具有一定意义。

3.2 结果计算与分析

本文利用 R 语言自编函数计算 V 指标和 Q 指标各数据项权重系数,结果如表 3、表 4 所示:

表 3 V 指标各数据项信息熵与权重系数

数据项	a_5	a_6	a_7	a_8	a_9	a_{10}
e	0.817 451	0.933 291	0.662 779	0.648 624	0.738 674	0.847 091
ω	0.135 013	0.049 338	0.249 407	0.259 876	0.193 276	0.113 091

表 4 Q 指标各数据项信息熵与权重系数

数据项	a_4	a_{11}	a_{12}	a_{13}	a_{14}
e	0.929 91	0.863 739	0.845 034	0.782 416	0.800 648
ω	0.090 061	0.175 085	0.199 12	0.279 58	0.256 153

熵权法基于数据的离散程度确定其权重,数据离散性越大所含的信息量越大,信息的不确定性越小,信息熵越小,相应地权重系数越大。从表 3 和表 4 中可以看出 V 指标和 Q 指标中对结果贡献最大的数据项分别是分享数和被推荐总数,贡献最小的数据项分别是活跃度数和头衔。上述现象表明,博客分享数和被推荐总数相较于其他数据项离散性大,包含的信息量最大且信息的不确定性最小,而活跃度数和头衔的数据特征与之相反。作为管理方的科学网一直鼓励用户生成、分享各类内容,将科学网博客打造成活跃的学术交流社区,上述指标权重分布特征可以更好地将持续生成、分享内容的活跃用户、权威用户识别出来。

基于上述权重系数,利用指标量化模型可分别计算博客积极性和权威性,部分结果如表 5、表 6 所示:

表 5 积极性指标 TOP20(Rank V)

排序	ID	V	排序	ID	V
1	41 174	0.212 805	11	69 474	0.077 082
2	281 238	0.190 206	12	69 051	0.069 937
3	558 553	0.158 403	13	235 687	0.069 414
4	39 946	0.148 529	14	469 915	0.068 064
5	91 121	0.129 378	15	2 277	0.064 922
6	107 667	0.105 898	16	1 352 130	0.064 394
7	350 729	0.104 976	17	1 557	0.057 538
8	280 034	0.097 539	18	43 310	0.056 212
9	433 662	0.095 634	19	1 750	0.055 006
10	415	0.088 333	20	69 474	0.077 082

表 6　权威性 TOP20(Rank Q)

排序	ID	Q	排序	ID	Q
1	1 557	0.311 567	11	2 277	0.101 708
2	41 174	0.201 156	12	69 474	0.101 396
3	53 483	0.165 464	13	254 303	0.101 351
4	415	0.148 427	14	279 992	0.100 941
5	41 757	0.148 189	15	2 237	0.097 328
6	40 247	0.124 394	16	55 745	0.092 425
7	280 034	0.122 326	17	111 635	0.086 453
8	117 889	0.111 436	18	2 984	0.085 491
9	575 129	0.108 149	19	39 731	0.082 146
10	279 177	0.107 498	20	4 699	0.080 555

受数据归一化影响,积极性指标和权威性指标值域在[0,1]之间。从计算结果看,两项指标整体取值水平较低。经过分析,上述现象的出现可主要归结为以下 3 个原因:①依据熵权法的特征,如果权重系数较大的数据项(如分享数和被推荐总数),用户行为数据整体表现不佳可能会造成指标取值水平偏低。②作为非正式学术社交平台,学术博客用户行为具有较强的随意性和不确定性。受到用户偏好和平台功能设置的影响,少部分用户单项指标缺失或表现突出,会造成指标取值水平偏低的情况。③实际使用过程中,只有少部分用户积极使用各项功能并持续为平台贡献高影响力博文。观察计算结果和计算活跃属性与权威属性的相关关系($r = 0.484$),发现部分积极性较大的用户其权威性也相对较大。该现象表明用户通过积极使用博客平台有助于提升权威性,拥有较高权威性的博客在博客平台相对活跃。基于 h 指数和 R 指数思想的博文影响力计算结果标准化值如表 7 所示:

表 7　博客博文影响力 TOP20(基于 hc 排序)

排序	ID	a_5	a_{15}	a_{13}	a_{16}	h_c	R_c	h_q	R_q
1	117 889	576	10 458 969	22 440	44 070	151	127.838 2	120	122.379 8
2	2 237	323	8 141 016	18 771	14 779	148	134.372 8	87	117.588 7
3	3 075	820	11 604 339	83 15	14 328	129	105.449 5	57	66.246 63
4	41 174	2 649	17 224 315	35 299	19 269	128	108.178 9	58	60.227 21
5	55 745	1 582	8 975 895	30 769	42 448	127	91.875 53	135	101.369 4
6	176	1 068	9 869 721	12 362	18 722	127	110.734 2	69	75.074 87
7	218 980	72 209	85 542 916	103 610	126 637	121	87.804 9	74	53.601 82
8	41 757	13 56	10 770 163	34 503	30 632	120	102.089 9	86	84.253 51
9	347 754	712	7 858 331	2 809	25 643	120	115.654 6	84	98.263 55
10	412 323	347	4 875 996	13 897	10 904	120	107.415 8	76	78.117 06
11	425 437	243	4 075 926	5 151	7 455	120	100.939 2	57	61.424 6
12	71 964	325	4 700 998	10 042	8 111	115	104.250 1	61	67.069 23
13	57 081	903	7 101 441	862	914	114	97.301 29	11	12.836 9
14	677 221	357	3 705 580	5 631	5 229	114	93.408 88	45	44.336
15	40 247	1 279	8 528 227	38 743	59 184	112	95.189 95	107	98.111 14

从表 7 计算结果看,h 指数克服了简单求和的数学逻辑缺陷,R 指数弥补了 h 指数取值相同时无法区分博文影响力的缺陷。R 指数作为补充说明指标,只需在 h 指数值相同时对比 R 指数值的大小。经过计算和观察,虽然 hc 指数与 hq 指数之间呈显著正相关关系($r = 0.743$),但两者仍存在一定差异,hc 指数高的博客其 hq 指数并不一定高。因此,将两者配对使用可以更全面地评估博客博文影响力。

本文以甲(ID = 1557)为例,生成系统－个人分类网,如图 1 所示。图 2 是基于 computeModules 函数的聚类结果,博客兴趣偏好以分类词组形式表示。

学术博客为非正式学术交流主要形式之一,从图 1 和图 2 可以看出,博客甲的博文内容按兴趣偏好分为 5 类,涵盖科研、科普、学习、教学和生活等方面,同时满足甲的学术和社交需求。加权结果按权重排序后,发现科学计量学研究和生活点滴类博文数量最多,其中科学计量学研究主要以博客资讯、观点评述、科研笔记、海外观察和论文交流等形式呈现。基于社区划分结果可提取分类词组"科学计量学、生活点滴"作为博客兴趣偏好标签。

基于上述研究,本文随机选择用户甲、乙、丙(ID 分别为 1 557、5 430、287 179)作为示例,利用用户画像模型 UPM 得到如表 8 所示的学术博客用户画像,表 9 为各指标评判阈值。

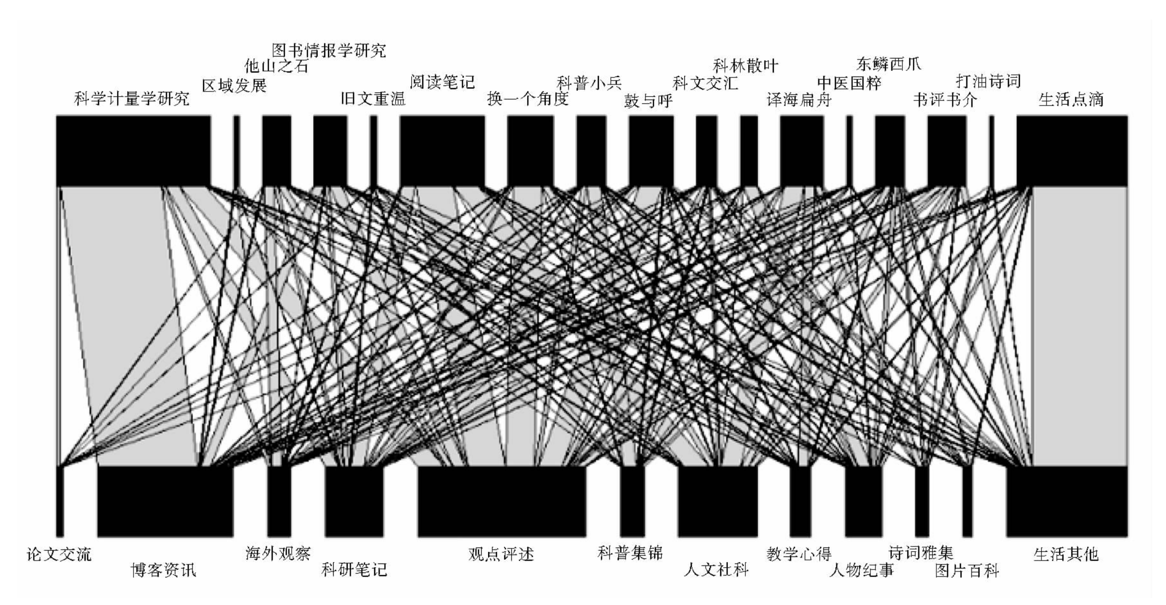


图 1　系统 – 个人分类二分网络

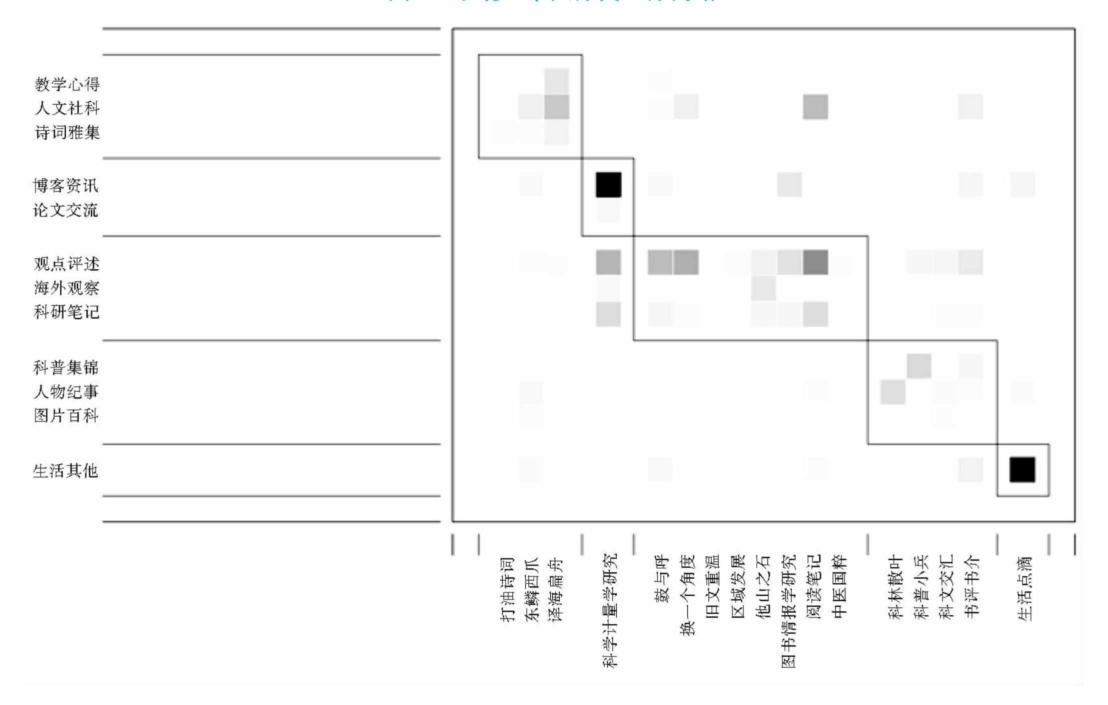


图 2　系统 – 个人分类网络的聚类

表 8　学术博客用户画像示例

维度		甲	乙	丙
基本属性	头衔	研究员	副教授	教授
	等级	5 级	4 级	5 级
	研究领域	管理综合	地球科学	信息科学
积极性	V	0.057 538	0.000 44	0.005 182
	等级	H	L	H
权威性	Q	0.311 567	0.003 292	0.040 225
	等级	H	C	H
影响力	阅读影响力	(107,92.27)	(52,42.12)	(108,94.22)
	等级	H	M	H
	互动影响力	(74,71.03)	(4,7.20)	(67,61.13)
	等级	H	L	H
兴趣偏好		科学计量学 生活点滴	海外观察 生活点滴	科研教学 科普札记

表 9　各指标评判阈值

等级	指标			
	V	Q	h_c	h_q
H	[0.212 805,0.004 327]	[0.311 567, 0.005 86]	[151,54]	[135,15]
M	[0.004 309,0.001 606]	[0.005 831,0.003 348]	[53,39]	[14,8]
C	[0.001 604,0.000 58]	[0.003346,0.002 014]	[38,21]	[7,5]

从表 8 中可以看出,研究领域的差异在一定程度影响博客的兴趣偏好,尤其是学术兴趣点。用户甲、丙通过在博客平台长期、持续贡献优质内容,积累了较高的权威性和较大的影响力,均可认为是科学网博客的优质用户,用户乙的表现则较为普通。上述用户画像

的结果可有多种用途,在精准营销情境下,平台管理员根据运营需求,结合用户画像的结果可识别出不同维度的高、中、普通等特征的用户并开展差异化营销,增强平台的核心竞争力;结合用户研究领域和兴趣偏好标签为各领域的用户尤其是新用户有针对性地推荐本领域的优质博客、博文,达到精准推荐和提升用户冷启动期间满意度的目的。用户还可以直接搜索感兴趣的标签来查找相关用户的信息资源,为好友关系的建立、学术交流、知识共享奠定基础,提高用户对学术博客平台服务的满意度。

4 结语

对学术博客为代表的在线学术社交平台开展用户画像研究具有一定的学术价值和现实意义。本文选择科学网博客用户行为数据为研究对象,以用户画像理论为基础,从基本属性、积极性、权威性、博文影响力和兴趣偏好 5 个维度构建了学术博客用户画像模型,并实际展示了具有代表性的用户画像示例。提出了标记学术博客用户特征的一些方法包括:①选择熵权法确定数据项权重系数;②从博文阅读和博文互动情况两个视角完善博文影响力评估指标体系,与正式文献交流的科学计量评价在方法上保持了一致性,利用 R 指数弥补 h 指数无法区分同值情况的不足;③基于系统分类与个人分类之间存在的二分网络关系,生成系统分类-个人分类加权二部图并划分社区,提取博客兴趣偏好标签。通过对博客用户画像,可以有效识别出平台的用户特征差异,服务平台的精准营销,提高冷启动时期的用户体验。由于数据集的限制,本文没有从博文内容抽取用户主题偏好,没能考虑时间对主题偏好和其他指标特征的影响。未来笔者会结合时域的概念对不同时间窗口下的用户行为数据特征,基于博文的内容提取用户主题偏好,得到更有意义的学术博客平台用户画像。

参考文献:

[1] COOPER A. About face 3 - the essentials of interaction design[C]//John Wiley & Sons,2007.
[2] 曾鸿,吴苏倪. 基于微博的大数据用户画像与精准营销[J]. 现代经济信息,2016(16):306-308.
[3] 冉蹬. 基于用户画像的手机游戏用户个性化内容推荐研究[D]. 西安:西安理工大学,2018.
[4] 李冰,王悦,刘永祥. 大数据环境下基于 K-means 的用户画像与智能推荐的应用[J]. 现代计算机(专业版),2016(24):11-15.
[5] 毕闰芳. 基于 SVR 的协同过滤与用户画像融合的电影个性化推荐研究[D]. 郑州:郑州大学,2018.
[6] 余孟杰. 产品研发中用户画像的数据模建——从具象到抽象[J]. 设计艺术研究,2014,4(6):60-64.
[7] XU G, ZHANG Y, ZHOU X. Towards user profiling for Web recommendation[J]. Lecture notes in computer science, 2005, 3809: 415-424.
[8] 张小可,沈文明,杜翠凤. 贝叶斯网络在用户画像构建中的研究[J]. 移动通信,2016,40(22):22-26.
[9] 周妹璇. 基于深度神经网络的用户画像研究[D]. 长沙:湖南大学,2018.
[10] 余传明,田鑫,郭亚静,等. 基于行为-内容融合模型的用户画像研究[J]. 图书情报工作,2018,62(13):54-63.
[11] 辛菊琴,蒋艳,舒少龙. 综合用户偏好模型和 BP 神经网络的个性化推荐[J]. 计算机工程与应用,2013,49(2):57-60.
[12] 马超. 基于主题模型的社交网络用户画像分析方法[D]. 合肥:中国科学技术大学,2017.
[13] 姚远. 基于本体的用户画像构建方法[C]//中国计算机用户协会网络应用分会. 第二十二届网络新技术与应用年会论文集. 北京:北京联合大学北京市信息服务工程重点实验室,2018.
[14] RAGHURAM M A, AKSHAY K, CHANDRASEKARAN K. Efficient user profiling in twitter social network using traditional classifiers[EB/OL]. [2019-05-20]. https://doi.org/10.1007/978-3-319-23258-4_35.
[15] JIAMTHAPTHAKSIN R, AUNG T H. User preferences profiling based on user behaviors on Facebook page categories[C]//International conference on knowledge & smart technology. Chonburi, Thailand: IEEE, 2017:248-253.
[16] 韩梅花,赵景秀. 基于用户画像的阅读疗法模式研究——以抑郁症为例[J]. 大学图书馆学报,2017,35(35):110.
[17] 王凌霄,沈卓,李艳. 社会化问答社区用户画像构建[J]. 情报理论与实践,2018(1):129-134.
[18] 刘海鸥,孙晶晶,张亚明,等. 在线社交活动中的用户画像及其信息传播行为研究[J]. 情报科学,2018,36(12):17-21.
[19] 崔超,罗欧. 科研知识社区中用户画像的实现思路[J]. 信息通信技术与政策,2018(6):75-78.
[20] 陈天歌. 基于社交媒体用户画像的品牌选择影响因素研究[D]. 广州:华南理工大学,2018.
[21] 周文静. 面向校园论坛用户兴趣的用户画像构建方法研究[D]. 北京:北京邮电大学,2018.
[22] 王琛. 学术博客影响力评价研究[D]. 太原:山西财经大学,2018.
[23] 郑超,陈峰. 科学家博客 h 指数与科学家 h 指数相关性分析[J]. 图书馆学研究,2013(3):53-57.
[24] 张晓阳,李晓亮. 科学家博客 h 指数评价及其相关性分析[J]. 图书情报工作,2010,54(2):66-69.

[25] 金碧辉,ROUSSEAU R. R指数、AR指数:h指数功能扩展的补充指标[J]. 科学观察,2007(3):1-8.

[26] 刘臣,吉莉,唐莉. 基于二分网中心节点识别的产品评论特征-观点词对提取研究[J]. 计算机系统应用,2018,27(11):9-16.

[27] 万国,张桂平,白宇,等. 基于特征加权的新闻主题句抽取[J]. 中文信息学报,2017,31(5):120-126.

作者贡献说明:

袁润:提出研究思路、设计并进行实验及论文修改;

王琦:采集实验数据、进行实验、论文撰写。

Construction and Empirical Study of User Portrait Model of Academic Blog: Taking ScienceNet as an Example

Yuan Run[1] Wang Qi[2]

[1] Library of Jiangsu University, Jiangsu. 212013

[2] Institute of Science and Technology Information, Jiangsu University, Jiangsu 212013

Abstract: [**Purpose/significance**] User portrait marks the behavioral characteristics of academic groups, which provides basis for user identification, precise marketing of academic social platform and improvement of user experience during cold boot period. [**Method/process**] The public users behavior data is obtained and processed by using Python and R language. The model of user portrait is constructed from five dimensions: user basic attribute, positivity, authority, blog post influence and interest preference. The empirical study takes the blog users behavior data of Science Web as an example. [**Result/conclusion**] This paper proposes specific indicators and calculation methods to characterize the user characteristics of academic blogs, which shows the user portrait model has certain theoretical significance and application value for the management and operation of academic social platforms.

Keywords: academic blog user portrait R language case analysis